PROCEEDINGS OF TDWG

Biodiversity Information Science and Standarts 1: e20275
doi: 10.3897/tdwgproceedings.1.20275

OPEN ACCESS

Conference Abstract

Defining a Data Quality (DQ) profile and DQ report using a prototype of Node.js module of the Fitness for Use Backbone (FFUB)

Allan K Veiga‡, Antonio M Saraiva‡

‡ Universidade de São Paulo, São Paulo, Brazil

Corresponding author: Allan K Veiga (allan.kv@gmail.com)

Received: 14 Aug 2017 | Published: 14 Aug 2017

Citation: Veiga A, Saraiva A (2017) Defining a Data Quality (DQ) profile and DQ report using a prototype of Node.js module of the Fitness for Use Backbone (FFUB). Proceedings of TDWG 1: e20275. https://doi.org/10.3897/tdwgproceedings.1.20275

Abstract

Despite the increasing availability of biodiversity data, determing the quality of data and informing would-be data consumers and users remains a significant issue. In order for data users and data owners to perform a satisfactory assessment and management of data fitness for use, they require a Data Quality (DQ) report, which presents a set of relevant DQ measures, validations, and amendments assigned to data.

Determining the meaning of "fitness for use" is essential to best manage and assess DQ. To tackle the problem, the TDWG Biodiversity Data Quality (BDQ) - Interest Group (IG) (https://github.com/tdwg/bdq) has proposed a conceptual framework that defines the necessary components to describe Data DQ needs, DQ solutions, and DQ reports (Fig. 1). It supports, in a global and collaborative environment, a consistent description of: (1) the meaning of data fitness for use in specific contexts, using the concept of a DQ profile; (2) DQ solutions, using the concepts of specifications and mechanisms; and (3) the status of quality of data according to a DQ profile, using the concept of a DQ report (Veiga 2016, Veiga et al. 2017). Based on this this conceptual framework, we implemented a prototype of a Fitness for Use Backbone (FFUB) as a Node.js module (https://nodejs.org/api/modules.html) for registering and retrieving instances of the framework concepts.

© Veiga A, Saraiva A. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

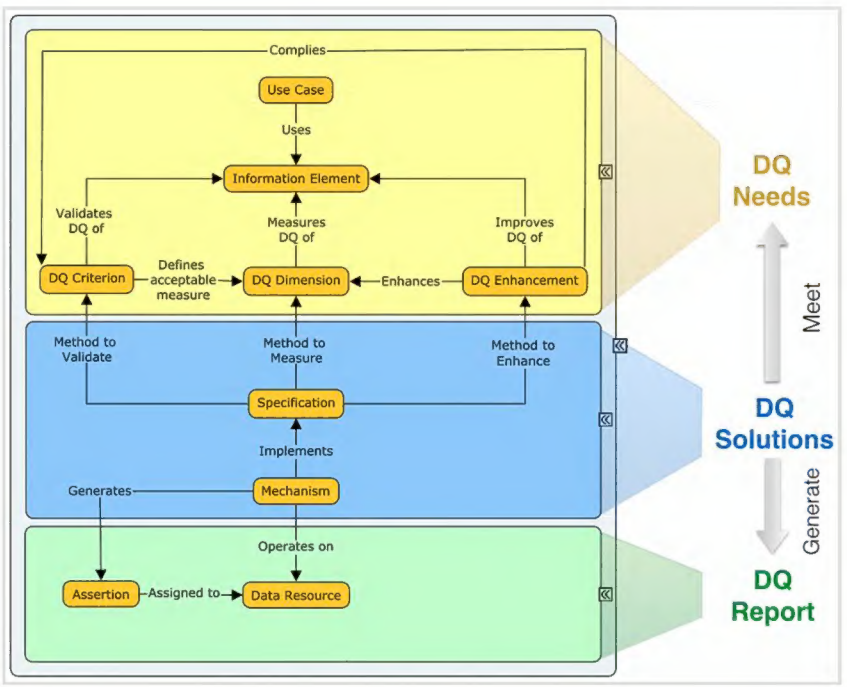

Figure 1.

The conceptual framework: Concepts and classes. DQ Needs concepts: Use Case, Information Element, DQ Dimension, DQ Criterion and DQ Enhancement. DQ Solutions concepts: Specification and Mechanism. DQ Report concepts: Data Source and Assertion (Veiga et al. 2017)

This prototype was built using Node.js, an asynchronous event-driven JavaScript runtime, which uses a non-blocking I/O model that makes it lightweight and efficient to build scalable network applications (https://nodejs.org). We registered our module in the npm package manager (https://www.npmjs.com) in order to facilitate its reuse and we made our source code available in GitHub (https://github.com) in order to foster collaborative development. To test the module, we developed a simple mechanism for measuring, validating and amending the quality of datasets and records, called BDQ-Toolkit, available in the FFUB module. The source code of the FFUB module can be found at https://github.com/BioComp-USP/ffub. Installing and using the module requires Node.js version 6 or higher. Instructions for installing and using the FFUB module can be found at https://www.npmjs.com/package/ffub (Veiga and Saraiva 2017).

Using the FFUB module we defined a simple DQ profile describing the meaning of data fitness for use in a specific context by registering a hypothetical use case. Then, we registered a set of valuable information elements for the context of the use case. For measuring the quality of each valuable information elements, we registered a set of DQ dimensions. To validate if the DQ measures are good enough, a set of DQ criteria was defined and registered. Lastly, a set of DQ enhancements for amending the quality in the use case context was also defined and registered. In order to describe the DQ solution used to meet those DQ needs, we registered the BDQ-Toolkit mechanism and all the specifications implemented by it. Using these specifications and mechanism, we generated and assigned to a dataset and its records a set of DQ assertions, according to the DQ dimensions, criteria and enhancements defined in the DQ profile. Based on those assertions we can build DQ reports by composing all the assertions assigned to the

dataset or to a specific record. This DQ report describes the status of DQ of a dataset or record according to the context of the DQ profile.

This module provides an interface to use the proposed conceptual framework, which allows others to register instances of its concepts. Future work will include creating a RESTful API using sophisticated methods of data retrieval.

Keywords

data quality, biodiversity informatics, fitness for use

Presenting author

Allan Koch Veiga

References

- Veiga AK (2016) A conceptual framework on biodiversity data quality. Ph.D. thesis. Universidade de São Paulo, São Paulo. Ph.D thesis URL: http://www.teses.usp.br/teses/disponiveis/3/3141/tde-17032017-085248/
- Veiga AK, Saraiva AM (2017) Toward a Biodiversity Data Fitness for Use Backbone (FFUB): a Node.js module prototype. TDWG 2017 Annual Conference, Ottawa, Canada. Proceedings of TDWG
- Veiga AK, Saraiva AM, Chapman AD, Morris PJ, Gendreau C, Schigel D, Robertson TJ (2017) A conceptual framework for quality assessment and management of biodiversity data. PloS one 12 (6): e0178731. https://doi.org/10.1371/journal.pone.0178731